Estatística básica em R

Vanderlei Júlio Debastiani (vanderleidebastiani@yahoo.com.br)

30 março 2019

Introdução

Este texto lista algumas funções em R utilizadas em estatística básica. Algumas funções relacionadas a medidas de posição e dispersão de variáveis aleatórias, testes de hipóteses, distribuições de probabilidade e modelagem estatística são apenas brevemente comentadas em tabelas. A ideia é que o leitor esteja ciente dos conceitos teóricos, aplicações e procedimentos utilizados ou ainda, que busque mais informações sobre os procedimentos e resultados retornados pelas funções.

Funções básicas

As funções básicas de estatística descritiva utilizadas em R estão listadas na Tabela 1 e alguns dos testes de hipóteses mais comuns estão listados na Tabela 2.

Tabela 1 – Funções básicas de estatística descritiva.

Função	**Descrição**
round	Arredondamento de valores
sqrt	Obter raiz quadrada
log	Obter logaritmo
exp	Obter exponencial
mean	Obter média
median	Obter mediana
var	Obter variância
sd	Obter desvio padrão
quantile	Obter quantis
cov	Obter covariância
cor	Obter correlação
table	Produzir tabelas de contingência

Tabela 2 – Algumas funções para aplicar testes de hipóteses.

Função	Teste
chisq.test	Teste qui-quadrado
t.test	Teste t
cor.test	Teste de correlação
shapiro.test	Teste de normalidade Shapiro
ncvTest, pacote car	Teste desvio não constante
TukeyHSD	Teste de comparação a posteriori de Tukey

Distribuições de probabilidade

Várias opções de distribuição de probabilidade estão disponíveis, cada uma possui quatro funções associadas, que são utilizadas para obter diferentes aspectos das distribuições. As descrições das funções estão detalhadas na Tabela 3 e algumas das distribuições na Tabela 4.

Tabela 3 – Funções associadas as distribuições de probabilidades. O nome das funções é composto pelo prefixo mais o nome da distribuição, conforme descrito na Tabela 4.

Prefixo	Descrição	Comentários
d___	Densidade probabilística ou densidade de massa	Obter a densidade probabilística ou de massa e a verossimilhança relativa de diferentes valores da variável aleatória
p___	Probabilidade acumulada	Obter probabilidade, dado valor da variável aleatória
q___	Quantil	Obter o valor associado da variável aleatória, dado uma probabilidade acumulada
r___	Simulação de valores	Amostrar valores da distribuição

Tabela 4 – Algumas distribuições de probabilidades disponíveis em R.

Distribuição	**Função base**
beta	dbeta
binomial (incluindo Bernoulli)	dbinom
Cauchy	dcauchy
qui-quadrado	dchisq
exponencial	dexp
F	df
gama	dgamma
geométrica	dgeom
hipergeométrica	dhyper
log-normal	dlnorm
multinomial	dmultinom
binomial negativa	dnbinom
Gaussiana (normal)	dnorm
Poisson	dpois
t de Student	dt
uniforme	dunif
Weibull	dweibull

Modelos estatísticos e fórmulas

A estrutura de qualquer modelo é especificada no R como um objeto da classe *formula*. A formatação é semelhante a utilizada em matemática.

$$variável\ resposta \sim variável\ explicativa$$

onde o símbolo do til (~) é lido como "é modelado como uma função de". Nos objetos da classe *formula* os símbolos aritméticos são usados de maneira diferente que os utilizados normalmente. Um resumo dos símbolos utilizados na atribuição das fórmulas pode ser encontrado na Tabela 5 e alguns exemplos na Tabela 6. As funções básicas para modelagem estatística estão listadas na Tabela 7 e algumas funções auxiliares na Tabela 8.

Tabela 5 – Símbolos utilizados na atribuição de objetos da classe *formula* no R. Traduzido com modificações de Crawley 2007.

Símbolos em fórmulas	**Comentários**
~	Indicar fórmula. Por exemplo y ~ x lido como "y em função de x"
+	Inclusão de uma variável explicativa no modelo
-	Indica exclusão de uma variável explicativa do modelo
*	Indica inclusão de variáveis explicativas e suas interações
/	Indica aninhamento de variáveis explicativas no modelo
|	Indica condicionamento. Por exemplo, y ~ x | z é lido como "y em função de x dado z"
I	Inibir a interpretação dos operadores "+", "-", "*", "/" e "^" nas fórmulas. São usados como operadores aritméticos

Tabela 6 – Alguns exemplos de fórmulas utilizadas no R. Traduzido com modificações de Crawley 2007.

Modelo	**Fórmula em R**	**Comentários**
Nulo	y ~ 1	1 é o intercepto nos modelos de regressão. Desta maneira é a média global
Regressão	y ~ x	x é uma variável explicativa contínua
Regressão através da origem	y ~ x-1	Não ajustar um intercepto
One-way ANOVA	y ~ sex	sex é uma variável categórica
One-way ANOVA	y ~ sex-1	Como acima, mas não se ajustar um intercepto (duas médias ao invés de uma média e uma diferença)
Two-way ANOVA	y ~ sex + genotype	genotype é uma variável categórica
ANOVA Fatorial	y ~ N*P*K	N, P e K são fatores. Todas as suas interações serão ajustadas
Three-way ANOVA	y ~ N*P*K – N:P:K	Como acima, mas não ajustar a interação de tripla
Análise de covariância	y ~ x + sex	Uma inclinação comum para y contra x, mas com um intercepto para cada nível da variável sex
Análise de covariância	y ~ x * sex	Uma inclinações e um intercepto para cada nível da variável sex
ANOVA aninhada	y ~ a/b/c	Fator c aninhado dentro do fator b e por sua vez aninhado dentro do fator a
ANOVA Split-plot	y ~ a*b*c+Error(a/b/c)	Um experimento fatorial com três diferentes variâncias de erro, uma para parcela
Regressão múltipla	y ~ x+z	Duas variáveis explicativas contínuas. Ajuste de superfície plana
Regressão múltipla	y ~ x*z	Ajustar um termo de interação (x + z + x: z)
Regressão múltipla	y ~ x + I(x^2) + z + I(z^2)	Ajustar um termo quadrático para x e z
Regressão múltipla	y ~ poly(x,2) + z	Ajustar um polinômio quadrático para x e linear para z
Regressão múltipla	y ~ (x+z+w)^2	Ajustar três variáveis e todas as suas interações até o polinômio de segundo grau
Modelo não paramétrico	y ~ s(x) +s(z)	y é uma função de x e z suavizados em um modelo aditivo generalizado
Transformação na variável no modelo	log(y) ~ I(1/x) + sqrt(z)	Todas as três variáveis são transformadas no modelo

Tabela 7 – Algumas funções de ajuste de modelos utilizadas no R. Traduzido com modificações de Crawley 2007.

Função	**Comentários**
lm	Ajustar um modelo linear com erros normais e variância constante. Geralmente é usado para análise de regressão usando variáveis explicativas contínuas
aov	Ajustar a análise de variância com erros normais, variância constante e função de ligação identidade. Geralmente usado para variáveis explicativas categóricas ou ANCOVA com uma combinação de variáveis explicativas categóricas e contínuas
glm	Ajustar modelos lineares generalizados usando dados de variáveis explicativas categóricas ou contínuas, especificando famílias de estruturas de erro (por exemplo, Poisson para dados de contagem ou binomial para dados binários e proporção) e outras função de ligação
nls	Ajustar modelos de regressão não lineares pelo método de mínimos quadrados e estimativa os parâmetros de funções não lineares
lmer, pacote lme4	Ajustar modelos lineares com efeitos fixos e aleatórios. Por exemplo, y~x+(1|grupo) para ajustar um intercepto para cada grupo e y~x+(x|grupo) um intercepto e uma inclinação para cada grupo

Tabela 8 – Algumas funções auxiliares para ajuste de modelos. Traduzido com modificações de Crawley 2007.

Funções auxiliares	**Comentários**
plot	Produz diagramas de diagnóstico para verificação de modelo, incluindo resíduos contra valores ajustados e testes de influência. Principais diagnósticos plot(modelo, which = c(1, 2, 4))
summary	Estimativas de parâmetros e erros padrão para função lm e tabelas ANOVA para função aov. As funções summary.aov ou summary.lm também podem ser usadas
anova	Comparar diferentes modelos e produzir tabelas ANOVA
resid ou residuals	Retornar resíduos
rstandard	Retornar resíduos normalizados
coef ou coefficients	Retornar os coeficientes (parâmetros estimados) do modelo
fitted	Retornar os valores ajustados e previstos pelo modelo. Apenas para as observações incluídas inicialmente no modelo
predict	Retornar os valores ajustados e previstos pelo modelo. Permite entrar com novos dados para calcular os valores previstos pelo modelo, intervalos de confiança, valores preditos na escala das variáveis preditoras ou na escala das funções de ligação. Geralmente usada para produzir linhas no gráfico de dispersão
points	Adicionar pontos (type = "p") ou linhas (type = "l") em qualquer gráfico. Geralmente usada em conjunto com a função predict
curve	Adicionar curvas a qualquer gráfico. Pode mostrar qualquer tipo de modelo, se fornecido uma expressão para calcular os valores da expressão em cada intervalo do gráfico
deviance	Retornar a deviância de um modelo
AIC	Retornar o AIC de um modelo
logLik	Retornar a log verossimilhança de um modelo
qqnorm e qqline	Gráfico QQ (quantil-quantil)
step	Seleção stepwise de modelos usando AIC
pchisq	Calcular p valor para distribuição qui-quadrado
update	Modificar o último modelo ajustado. Usado para remover ou adicionar variáveis aos modelos
mle2, pacote bbmle	Estimar parâmetros pelo método de máxima verossimilhança
Ictab, pacote bbmle	Gerar tabelas de AIC, AICc e outros critérios de informação

Conclusão

O objetivo deste texto foi apenas listar algumas funções básicas de estatística na utilizadas na linguagem R. Os principais símbolos utilizados para a definição de fórmulas e alguns exemplos foram comentados, algumas funções relacionadas a distribuições de probabilidade e modelagem estatística foram brevemente exemplificadas. Espero que este texto tenha sido útil e, por favor, avise-me se tiver dúvidas ou sugestões sobre este texto.

Mais informações

Outros textos e tutoriais sobre R podem ser encontrados em https://vanderleidebastiani.github.io/tutoriais.

Referências

Crawley, Michael J. 2007. **The R book**. John Wiley & Sons, Chichester.

R Core Team; 2018. **R Language Definition**. https://cran.r-project.org/doc/manuals/R-lang.html